Computerized adaptive testing for the random weights linear logistic test model

Marjolein Crabbe

Martina Vandebroek

Abstract

This paper discusses four item selection rules to design efficient individualized tests for the random weights linear logistic test model: minimum posterior weighted $\mathcal{D}$-error ($\mathcal{D}_B$), minimum expected posterior weighted $\mathcal{D}$-error ($E\mathcal{D}_B$), maximum expected Kullback–Leibler divergence between subsequent posteriors (KLP), and maximum mutual information (MUI). The random weights linear logistic test model decomposes test items into a set of subtasks or cognitive features and assumes individual-specific effects of the features on the difficulty of the items. The model extends and improves the well known linear logistic test model in which feature effects are only estimated at the aggregate level. Simulations show that the efficiencies of the designs obtained with the different criteria appear to be equivalent. However, KLP and MUI are given preference over $\mathcal{D}_B$ and $E\mathcal{D}_B$ due to their lesser complexity, which significantly reduces the computational burden.

Keywords: random weights linear logistic test model, computerized adaptive testing, $\mathcal{D}$-efficiency, Kullback–Leibler divergence, mutual information

1 Introduction

The linear logistic test model (LLTM) achieved great popularity and value in item response theory (IRT) for item complexity modeling. Since its introduction by Fischer in 1973, the LLTM has been applied in various educational and psychological settings to analyze the cognitive structure of test items, more specifically the decomposition of test items into a specific set of item stimulus features or subtasks and the influence of these features on the difficulty of the items (Bouwmeester et al., 2011; Embretson and Daniel, 2008; Freund et al., 2008; Holling et al., 2009; Hornke and Habon, 1986; Medina-Diaz, 1993; Poinstingl, 2009; Zeuch et al., 2011). But despite the LLTM's ease of interpretation and its many successful and instructive applications, one might question its assumption of fixed feature effects. Differing abilities might cause the presence of specific subtasks in an item to differently affect different people (and their probability of correctly solving the item). Because of these possible differences between individuals in the difficulty of a subtask, Rijmen and De Boeck (2002) extended the LLTM to the random weights linear logistic test model (RWLLTM), which allows for individual feature effects and therefore allows for potential heterogeneity in the aptitude of the test takers or, equivalently, in the individuals' abilities.

Like Cognitive Diagnostic Models (CDM) (Rupp et al., 2010), the RWLLTM makes use of a Q-matrix with attributes (features) so that individual differences in the performance of subjects can be related to these attributes. While the CDM approach reduces the individual differences to two or a small number of mastery levels per attribute and makes commonly use of binary attributes and more recently also of ordered category attributes, the RWLLTM can easily handle all kinds of attributes, including continuous attributes, and the individual differences with respect to these attributes are continuous instead of categorical. It is therefore a useful approach for diagnostic purposes, in an educational measurement context as well as in broader assessment contexts, such as personality assessment as will be illustrated with a data example.

The efficient design of tests for the RWLLTM has however not yet been addressed in the test design literature. As individual coefficients are present in the model, this research advo-

cates an individualized design approach. Four different item selection rules are compared to construct individualized tests by means of computerized adaptive testing (CAT). The tests are sequentially designed for each person separately, using prior information on the parameters and one's responses to previous test items. As such, the tests are tailored to the specific abilities of an individual, resulting in higher quality test data and in turn in more accurate parameter estimates. The first two design algorithms apply the well known $\mathcal{D}$-efficiency criterion and maximize the determinant of the model's Fisher information matrix. Note that one criterion uses Bayesian updates of the posterior distribution of the random effects to weigh the $\mathcal{D}$-criterion (van der Linden, 1998; van der Linden and Pashley, 2010), whereas the other considers expected updates of posteriors for weighing (van der Linden, 1998; van der Linden and Pashley, 2010). In contrast, the third and the fourth item selection rule are based on Kullback–Leibler divergence between subsequent posterior distributions (Mulder and van der Linden, 2010; Wang and Chang, 2011). Besides comparing the criteria with each other, they are also evaluated against the random selection of test items.

The criteria above are not new in adaptive IRT design and have already been applied to a couple of item response models. The present paper is, however, the first to apply them in the design of individualized tests for the RWLLTM. Van der Linden (1998) and van der Linden and Pashley (2010), for example, compare the design criteria based on $\mathcal{D}$-efficiency for, respectively, the two- and the three-parameter logistic model. Wang and Chang (2011), on the other hand, compare both the Fisher information and the Kullback–Leibler information item selection criteria for the multidimensional three-parameter logistic model. Note that instead of weighting $\mathcal{D}$-errors over posteriors, the Fisher information is evaluated at posterior modes in their paper, yielding only locally efficient designs. They found an improved estimation accuracy with the Kullback–Leibler criterion. As the RWLLTM is both conceptually and analytically very different from the models in those papers, this study provides an instructive extension in the research on individualized test design.

In the next section, the random weights linear logistic test model and the four item selection rules are introduced. The main study comparing the design criteria in constructing efficient

individualized tests for the RWLLTM is discussed in Section 3. The results as to the estimation accuracy, item exposure, item overlap, and computation time are given. The final section discusses the key findings.

2 Method

2.1 The random weights linear logistic test model

Fischer's (1973) linear logistic test model (LLTM) assumes that test items can be decomposed into a set of subtasks, also referred to as cognitive operations, item rules, attributes, or features. For each item in a test it can then be indicated whether or not a specific subtask must be completed in order to get the right answer. Consequently, the difficulty of an item is defined as the weighted sum of the required features, with weights expressing the relative effect of a subtask on the difficulty of the item. In the LLTM, the probability that person i solves item j correctly is modeled by (Fischer, 1973)

$$P(Y_{ij} = 1|\mathbf{x}_j; \phi_i, \boldsymbol{\beta}) = \frac{\exp(\phi_i + \sum_m \beta_m x_{jm})}{1 + \exp(\phi_i + \sum_m \beta_m x_{jm})}, \tag{1}$$

with ϕ_i the ability parameter for person i (measuring overall proficiency), x_{jm} the score of item j on item feature m, which is 1 if subtask m is required to solve item j correctly and 0 otherwise, and β_m the weight corresponding to feature m's relative effect on the difficulty of the item. Once the cognitive structure of the test items has been determined by the LLTM, one is able to construct (infinitely many) items with a specific set of features, and hence with a specific level of difficulty. Note that the Rasch model is essentially a special case of the LLTM, assuming each item in a test represents a subtask. Instead of estimating the difficulty weights for the item features, an item difficulty parameter is estimated for each item separately.

The fixed effects assumption of the item features on the item difficulty in (1) might, however, not always be realistic. An intuitive example is that of mathematical questions for which the problem is described in a short text and thus requires both mathematical and verbal skills

for their solve (Rijmen and De Boeck, 2002). As not all persons have the same analytical or linguistic competence, it is no surprise that heterogeneity in feature effects may arise. To take this into account, the random weights linear logistic test model (RWLLTM) extends the LLTM and assumes individual effects on the probability of success for (some or all of) the item features. The random effects measure the differences in subtask difficulty between individuals and therefore differences in the individuals' abilities.

The total effect of an item feature s that is assumed to have an individual effect is denoted by the mean effect β_s and an individual deviation from this mean, denoted by θ_{is}. For those random features, the item scores are duplicated in the variables z. The RWLLTM probability that person i passes item j is (Rijmen and De Boeck, 2002)

$$P(Y_{ij} = 1|\mathbf{x}_j; \phi_i, \boldsymbol{\beta}, \boldsymbol{\theta}_i) = \frac{\exp(\phi_i + \sum_m \beta_m x_{jm} + \sum_s \theta_{is} z_{js})}{1 + \exp(\phi_i + \sum_m \beta_m x_{jm} + \sum_s \theta_{is} z_{js})} \quad (2)$$

$$= \frac{\exp(\phi_i + \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\theta}_i)}{1 + \exp(\phi_i + \mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\theta}_i)}, \quad (3)$$

with $\mathbf{x}_j$ including the binary item scores for all features considered and $\mathbf{z}_j$ including only the scores for the item features with a random effect. Note that the coefficient ϕ_i measures the ability not specified by the item features. As this overall ability is also random, it can be incorporated in the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_i$. More specifically, the mean of ϕ_i is included in $\boldsymbol{\beta}$, whereas the individual part is included in $\boldsymbol{\theta}_i$. The response probability in the RWLLTM thus becomes

$$P(Y_{ij} = 1|\mathbf{x}_j; \boldsymbol{\beta}, \boldsymbol{\theta}_i) = \frac{\exp(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\theta}_i)}{1 + \exp(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\theta}_i)}, \quad (4)$$

with the first element in both $\mathbf{x}_j$ and $\mathbf{z}_j$ equal to 1 for every item (to represent the ability of each person). The vector $\boldsymbol{\beta}$ thus contains the coefficients for the features with a fixed effect and the average effect of the features with individual effects. In the remainder of the present paper, we will assume $\mathbf{x}_j$ is p-dimensional and $\mathbf{z}_j$ a $q \times 1$ subset of it, with q the number of parameters corresponding to individual effects. Further, the model assumes a heterogeneity distribution,

in most cases a multivariate normal distribution, over the individual-specific coefficients

$$\boldsymbol{\theta}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}). \tag{5}$$

Therefore, the marginal likelihood of the RWLLTM for a sample of N test takers and a test including J items is

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{D}|\mathbf{Y}, \mathbf{X}) &= \prod_{i=1}^{N} \int L(\boldsymbol{\beta}, \boldsymbol{\theta}_i|\mathbf{y}_i, \mathbf{X})\ \phi(\boldsymbol{\theta}_i|\mathbf{0}, \mathbf{D})\ d\boldsymbol{\theta}_i & (6) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{J} \frac{\exp[y_{ij}(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\theta}_i)]}{1 + \exp(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\theta}_i)}\ \phi(\boldsymbol{\theta}_i|\mathbf{0}, \mathbf{D})\ d\boldsymbol{\theta}_i, & (7) \end{aligned}$$

with $\mathbf{y}_i = (y_{i1}, ..., y_{iJ})'$ the response vector of person i and $y_{ij} = 1$ if person i correctly solves item j, and 0 otherwise, $\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_J')'$ the test design, and ϕ the normal density. Note that the design matrix $\mathbf{X}$ is equivalent to the $\mathbf{Q}$ matrix in cognitive diagnosis models.

Rijmen and De Boeck (2002) illustrate the benefits of the RWLLTM by means of a dataset on deductive reasoning. We will use a data set on verbal aggression from De Boeck et al. (2011), to show that including more individualized effects can indeed increase significantly the model fit, and to clarify the notation used. The verbal aggression data set consists of the responses of 316 test takers to 24 items on verbal aggression. All items represent a frustrating situation and the subjects were asked whether or not they would or would want to react in a verbally aggressive way. The possible verbal aggressive responses are: cursing, scolding, and shouting. Note that this experiment tests an attitude instead of abilities, nevertheless the RWLLTM can be properly applied. All items in the test are characterized by four dummy variables, i.e., one to indicate whether the situation is self-to-blame or other-to-blame, two for representing the behavior type (cursing, scolding, or shouting), and one giving the mode of the item (wanting or doing). In addition, the intercept represents the average verbal aggressiveness for all items in the test. Consequently, the vector $\boldsymbol{\beta}$ is five dimensional. Assuming effects coding, the vector $\mathbf{x}_j = (1\ -1\ 1\ 0\ 1)'$ then corresponds to an observation for which the situation is *other-to-blame*

Table 1: *Log-likelihood (LL), AIC and BIC Values for Different RWLLTMs*

Random effects	q	LL	AIC	BIC
Intercept + Type + Mode + Situation	5	-3903	7847	7985
Intercept + Type + Mode	4	-3951	7933	8037
Intercept + Type + Situation	4	-3961	7953	8057
Intercept + Mode + Situation	3	-4043	8108	8185
Intercept + Type	3	-4003	8029	8105
Intercept + Mode	2	-4081	8178	8233
Intercept + Situation	2	-4086	8188	8243
Intercept (LLTM)	1	-4119	8250	8292

(second element), the behavior type is *cursing* (third and fourth element) and the mode equals *wanting* (fifth element). Remember that the first element in $\mathbf{x}_j$ stands for the intercept.

Models with some or all of the features having an individual effect were estimated. These models, their log-likelihood (LL), AIC, and BIC values were obtained with the `lmer` function and are given in Table 1.

For example, the second model in the table assumes that, except for the general verbal aggressiveness, only type and mode have individual effects. Therefore the vector $\mathbf{z}_j$ corresponding to $\mathbf{x}_j = (1 \ -1 \ 1 \ 0 \ 1)'$ is $\mathbf{z}_j = (1 \ 1 \ 0 \ 1)'$, retaining only the elements corresponding to the intercept, type and mode.

It is clearly beneficial to consider random coefficients for the features characterizing the test items. The AIC and BIC values decrease for models with more individual effects. The RWLLTM with all feature effects random even yields the lowest values of AIC and BIC. The improvement in the model fit from letting the features have random effects is significant, as can be derived from likelihood ratio tests. The test yields p-values smaller than 0.01% for every pairwise comparison of nested models.

2.2 Computerized adaptive testing for the RWLLTM

An important objective of modeling a RWLLTM is the accurate estimation of the random effects $\boldsymbol{\theta}_i$. Therefore, instead of designing a fixed test for all examinees, it is more sensible and efficient to construct individualized tests customized to the specific abilities of the examinee,

improving the quality of the test data. The next sections describe four different item selection rules to construct efficient individualized test designs for the RWLLTM. The methods will be compared in a simulation study as to their design efficiency, item exposure, item overlap, and computation time.

2.2.1 Minimum posterior weighted $\mathcal{D}$-error

The $\mathcal{D}$-efficiency criterion (Atkinson et al., 2007) is a well known and widely used measure for constructing efficient designs. In a multidimensional setting, this criterion equals the determinant of the model's Fisher information matrix, which is the negative expectation of the second order partial derivative of the log-likelihood function. The rationale behind this criterion is that the determinant of the Fisher information matrix is inversely proportional to the volume of the confidence ellipsoid around the maximum likelihood parameter estimates. Therefore, an efficient test design is a design that maximizes the $\mathcal{D}$-efficiency criterion. Note that here, as Bayesian inference is assumed for the individual models, the logarithm of the posterior instead of the log-likelihood is incorporated in the criterion (Mulder and van der Linden, 2009; Segall, 2010; Yu et al., 2011) yielding a Bayesian Fisher information matrix (BFIM).

The information matrix for a test design $\mathbf{X}$ with respect to the random coefficients $\boldsymbol{\theta}_i$ of an individual is then given by

$$\mathbf{I}_{BFIM}(\boldsymbol{\theta}_i, \mathbf{X}) = -\mathrm{E}\left[\frac{\partial^2 \log[L(\boldsymbol{\beta}, \boldsymbol{\theta}_i|\mathbf{y}, \mathbf{X}) f_0(\boldsymbol{\theta}_i)]}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i'}\right], \tag{8}$$

with $f_0(\boldsymbol{\theta}_i)$ a prior distribution for the random feature effects. Assuming prior values $\boldsymbol{\beta}_0$ for the mean effects and $f_0(\boldsymbol{\theta}_i) \equiv \phi(\boldsymbol{\theta}_i|\mathbf{0}, \mathbf{D}_0)$, the normal density with variance covariance matrix $\mathbf{D}_0$, one can show that

$$\mathbf{I}_{BFIM}(\boldsymbol{\theta}_i, \mathbf{X}) = \sum_{j=1}^{J} \frac{\exp(\mathbf{x}_j'\boldsymbol{\beta}_0 + \mathbf{z}_j'\boldsymbol{\theta}_i)}{[1 + \exp(\mathbf{x}_j'\boldsymbol{\beta}_0 + \mathbf{z}_j'\boldsymbol{\theta}_i)]^2} \mathbf{z}_j \mathbf{z}_j' + \mathbf{D}_0^{-1}. \tag{9}$$

Equivalent to maximizing the determinant of $\mathbf{I}_{BFIM}$, we will minimize the inverse, referred to as the $\mathcal{D}$-error,

$$\mathcal{D} = \det[\mathbf{I}_{BFIM}(\boldsymbol{\theta}_i, \mathbf{X})]^{-1/q}. \tag{10}$$

Moreover, instead of fixing the random effects $\boldsymbol{\theta}_i$ in the criterion to some specific prior values or estimates and therefore obtaining merely local optimality, the $\mathcal{D}$-error is averaged over a prior distribution of the random effects, which is updated each time an item response is observed.

As no answers are available at the beginning of the test, the Bayesian $\mathcal{D}$-error is minimized over all possible test items with respect to a preset prior distribution $f(\boldsymbol{\theta}_i)$ of the random effects, obtaining the first item

$$\mathcal{D}_B = \int \det[\mathbf{I}_{BFIM}(\boldsymbol{\theta}_i, \mathbf{x}_1)]^{-1/q} \; f(\boldsymbol{\theta}_i) \; d\boldsymbol{\theta}_i. \tag{11}$$

For this and all subsequent design criteria, we set the initial prior $f(\boldsymbol{\theta}_i)$ to $\phi(\boldsymbol{\theta}_i|\mathbf{0}, \mathbf{D}_0)$. Note that although here this "design" prior equals the "inference" prior $f_0(\boldsymbol{\theta}_i)$ in (9), this is not necessary. After $k-1$ test items have been solved by the examinee, the prior information on $\boldsymbol{\theta}_i$ can be updated with the observed response data in the following way

$$f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \;\; = \;\; \frac{L(\boldsymbol{\beta}_0, \boldsymbol{\theta}_i|\mathbf{y}_{k-1}, \mathbf{X}_{k-1}) \; \phi(\boldsymbol{\theta}_i|\mathbf{0}, \mathbf{D}_0)}{\int L(\boldsymbol{\beta}_0, \boldsymbol{\theta}_i|\mathbf{y}_{k-1}, \mathbf{X}_{k-1}) \; \phi(\boldsymbol{\theta}_i|\mathbf{0}, \mathbf{D}_0) \; d\boldsymbol{\theta}_i}, \tag{12}$$

with $\mathbf{y}_{k-1} = (y_1, ..., y_{k-1})'$. The kth item in the test is now obtained by minimizing the Bayesian $\mathcal{D}$-error (for the design $\mathbf{X}_k$ including the $k-1$ previously administered test items and the current candidate for item k), weighted over this posterior distribution:

$$\mathcal{D}_B = \int \det[\mathbf{I}_{BFIM}(\boldsymbol{\theta}_i, \mathbf{X}_k)]^{-1/q} \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \; d\boldsymbol{\theta}_i. \tag{13}$$

2.2.2 Minimum expected posterior weighted $\mathcal{D}$-error

The above design algorithm updates the posterior distribution of the random effects with the observed responses from the previous test items. The current criterion, in contrast, updates the posteriors up until the candidate item, averaging over all possible responses to that item.

As only dichotomous test items are assumed here, the answer is either correct (1) or incorrect (0). The kth test item is thus found by minimizing the expected posterior weighted $\mathcal{D}$-error (van der Linden, 1998; van der Linden and Pashley, 2010), which is given by

$$E\mathcal{D}_B = \sum_{y_k=0}^{1} f(y_k|\mathbf{y}_{k-1}) \int \det[\mathbf{I}_{BFIM}(\boldsymbol{\theta}_i, \mathbf{X}_k)]^{-1/q} \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k) \; d\boldsymbol{\theta}_i, \tag{14}$$

with

$$\begin{aligned} f(y_k|\mathbf{y}_{k-1}) &= \int f(y_k|\boldsymbol{\theta}_i) \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \; d\boldsymbol{\theta}_i & (15)\\ f(y_k|\boldsymbol{\theta}_i) &= p_k^{y_k}(1-p_k)^{1-y_k} & (16)\\ p_k &= P(Y_k = 1|\mathbf{x}_k; \boldsymbol{\beta}_0, \boldsymbol{\theta}_i). & (17) \end{aligned}$$

Van der Linden (1998) and van der Linden and Pashley (2010) label the weights $f(y_k|\mathbf{y}_{k-1}), y_k = 0, 1$ in (14) as the posterior predictive probability function.

2.2.3 Maximum expected Kullback–Leibler divergence between subsequent posteriors

Chang and Ying (1996) introduced a totally different approach for optimal item selection in individualized test design, by using Kullback–Leibler information to distinguish between test items. The Kullback–Leibler divergence between two densities f and g for a continuous variable X is (Mulder and van der Linden, 2010)

$$\begin{aligned} KL(f,g) &= \mathrm{E}_f\left[\log\frac{f(x)}{g(x)}\right] & (18)\\ &= \int f(x) \; \log\frac{f(x)}{g(x)} \; dx. & (19) \end{aligned}$$

For any two f and g, $KL(f,g) \geq 0$, and $KL(f,f) = 0$, and hence the Kullback–Leibler divergence is commonly interpreted as the distance between the two densities. Note, however, that in contrast to real distance measures, the Kullback–Leibler divergence is not symmetric.

Mulder and van der Linden (2010) extended the ideas of Chang and Ying (1996) and applied

the Kullback–Leibler divergence to the subsequent posteriors of the individual coefficients in an item response model. They argued that as the subsequent item in a test should extract as much additional information as possible, it should maximize the distance between the current posterior distribution of the individual coefficients and the updated posterior one obtains with the response to the next item. As with the expected posterior weighted $\mathcal{D}$-error, they average the Kullback–Leibler divergence between posteriors over the possible responses to the candidate item, weighting both options (correct or incorrect) with the posterior predictive probability function. The kth item is thus found by maximizing the expected Kullback–Leibler distance between subsequent posteriors, i.e.,

$$KLP = \sum_{y_k=0}^{1} f(y_k|\mathbf{y}_{k-1})\ KL[f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}), f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k)]. \tag{20}$$

Note that this criterion can be rewritten as (Mulder and van der Linden, 2010)

$$\begin{aligned} KLP &= f(0|\mathbf{y}_{k-1})\ \log\ f(0|\mathbf{y}_{k-1}) + f(1|\mathbf{y}_{k-1})\ \log\ f(1|\mathbf{y}_{k-1}) \\ &\quad -f(0|\mathbf{y}_{k-1}) \int \log\ (1-p_k)\ f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1})\ d\boldsymbol{\theta}_i - f(1|\mathbf{y}_{k-1}) \int \log\ p_k\ f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1})\ d\boldsymbol{\theta}_i. \end{aligned} \tag{21}$$

2.2.4 Maximum mutual information

Related to the Kullback–Leibler divergence is the mutual information between two variables X and Y, defined as (Mulder and van der Linden, 2010)

$$I_M(X,Y) = \int_Y \int_X f(x,y)\ \log \frac{f(x,y)}{f(x)f(y)}\ dxdy. \tag{22}$$

It is the Kullback–Leibler distance between the joint distribution of X and Y and their distribution in case of independence. It measures how much information about each other is captured by the variables.

Transforming this into an item selection criterion, Mulder and van der Linden (2010) and Wang and Chang (2011) suggest maximizing the mutual information between the posterior distribution of the random effects and the distribution of the response to the candidate item,

given the previous responses,

$$
\begin{aligned}
MUI &= \sum_{y_k=0}^{1} \int f(\boldsymbol{\theta}_i, y_k|\mathbf{y}_{k-1}) \log \frac{f(\boldsymbol{\theta}_i, y_k|\mathbf{y}_{k-1})}{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) f(y_k|\mathbf{y}_{k-1})} \, d\boldsymbol{\theta}_i && (23)\\
&= \sum_{y_k=0}^{1} \int f(y_k|\boldsymbol{\theta}_i) \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \log \frac{f(y_k|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1})}{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) f(y_k|\mathbf{y}_{k-1})} \, d\boldsymbol{\theta}_i && (24)\\
&= \sum_{y_k=0}^{1} \int f(y_k|\boldsymbol{\theta}_i) \log \frac{f(y_k|\boldsymbol{\theta}_i)}{f(y_k|\mathbf{y}_{k-1})} \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \, d\boldsymbol{\theta}_i && (25)\\
&= \int (1-p_k) \log (1-p_k) \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \, d\boldsymbol{\theta}_i + \int p_k \log p_k \; f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) \, d\boldsymbol{\theta}_i - \\
&\quad f(0|\mathbf{y}_{k-1}) \log f(0|\mathbf{y}_{k-1}) - f(1|\mathbf{y}_{k-1}) \log f(1|\mathbf{y}_{k-1}). && (26)
\end{aligned}
$$

The relation between the mutual information and the Kullback–Leibler distance becomes even more clear from the following calculations, proving that the mutual information criterion is in essence the expected Kullback–Leibler distance between the updated and the current posterior (Mulder and van der Linden, 2010; Wang and Chang, 2011):

$$
\begin{aligned}
MUI &= \sum_{y_k=0}^{1} \int f(\boldsymbol{\theta}_i, y_k|\mathbf{y}_{k-1}) \log \frac{f(\boldsymbol{\theta}_i, y_k|\mathbf{y}_{k-1})}{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) f(y_k|\mathbf{y}_{k-1})} \, d\boldsymbol{\theta}_i && (27)\\
&= \sum_{y_k=0}^{1} \int f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k) \; f(y_k|\mathbf{y}_{k-1}) \log \frac{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k) f(y_k|\mathbf{y}_{k-1})}{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}) f(y_k|\mathbf{y}_{k-1})} \, d\boldsymbol{\theta}_i && (28)\\
&= \sum_{y_k=0}^{1} f(y_k|\mathbf{y}_{k-1}) \int f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k) \log \frac{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k)}{f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1})} \, d\boldsymbol{\theta}_i && (29)\\
&= \sum_{y_k=0}^{1} f(y_k|\mathbf{y}_{k-1}) \; KL[f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1}, y_k), f(\boldsymbol{\theta}_i|\mathbf{y}_{k-1})]. && (30)
\end{aligned}
$$

Although KLP and MUI are thus very similar (only the order of the posteriors in (20) and (30) is reversed), they are not equal, as the Kullback–Leibler measure is non-symmetric. Mulder and van der Linden (2010) and Wang and Chang (2011) apply the Kullback–Leibler criterion to the multidimensional three-parameter logistic model. The present paper, however, explores this criterion for the RWLLTM.

Note that the criteria in (21) and (26) only involve the posterior weighted response probabilities for the candidate item, whereas the $\mathcal{D}_B$ and $E\mathcal{D}_B$ criteria require the computation of

the $\mathcal{D}$-error, including also all previously administered items in the test. Clearly, $\mathcal{D}_B$ and $E\mathcal{D}_B$ are computationally more intensive than KLP and MUI, which negatively affects their computation time. In practice, the integrals in the efficiency criteria are approximated by averages over draws from the distribution at hand. In case of a normal prior in the beginning of the test, draws are easily obtained. Unfortunately, the updated posteriors after each observed response do not have a closed form, which complicates the sampling. Therefore, importance sampling is used to approximate the integrals (Yu et al., 2011). The details of this approximation technique are given in Appendix A which is available online. Other approaches, such as normal approximation (Mulder and van der Linden, 2010) or Gauss–Hermite quadrature (Wang et al., 2011), might also be applied.

3 Comparison study of the design criteria

In this section, minimum posterior weighted $\mathcal{D}$-error ($\mathcal{D}_B$), minimum expected posterior weighted $\mathcal{D}$-error ($E\mathcal{D}_B$), maximum expected Kullback–Leibler divergence between subsequent posteriors (KLP) and maximum mutual information (MUI) are compared as item selection rule to construct individual sequential designs for the RWLLTM and are also evaluated against the random selection of test items.

To measure the accuracy of the estimates obtained with the different individual designs, the root mean squared errors (RMSE) are computed. They are a measure of the total estimation error and are computed separately for all $\boldsymbol{\theta}_i$, $\boldsymbol{\beta}$ and $\mathbf{D}$. For the individual effects $\boldsymbol{\theta}_i$, the RMSE equals

$$\text{RMSE}_{\boldsymbol{\theta}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)}{q}}, \tag{31}$$

with $\hat{\boldsymbol{\theta}}_i$ the estimates of person i's individual coefficients, $\boldsymbol{\theta}_i$ the true (simulated) values for person i's individual coefficients, N the number of test takers, and q the number of random coefficients in the model. Although the main focus of this paper is on the precise estimation of the individual ability and feature effects, we will also have a look at the precision of the estimated fixed and mean random effects $\boldsymbol{\beta}$ and at the matrix $\mathbf{D}$ which represents tha heterogeneity in

the population and the covariances among the feature effects. Define

$$\text{RMSE}_{\boldsymbol{\beta}} = \sqrt{\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p}}, \tag{32}$$

with $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ respectively the estimated and true fixed and mean effects and p the total number of coefficients.

Finally, to compute the RMSE for $\mathbf{D}$, the variance covariance matrix is transformed into a vector $\mathbf{d}$ containing all distinct elements of the matrix. $\text{RMSE}_{\mathbf{D}}$ is given by

$$\text{RMSE}_{\mathbf{D}} = \sqrt{\frac{(\hat{\mathbf{d}} - \mathbf{d})'(\hat{\mathbf{d}} - \mathbf{d})}{q_d}}, \tag{33}$$

with $\hat{\mathbf{d}}$ the estimates, $\mathbf{d}$ the corresponding true coefficients, and q_d the number of distinct elements in the variance covariance matrix, i.e., $q(q+1)/2$. Note that MCMC estimation, more specifically Gibbs sampling, is used in this simulation study to estimate the models, as several studies have already illustrated the high estimation and prediction precision obtained from this approach (see for instance Allenby et al., 1995; Arora et al., 1998; Arora and Huber, 2001). Some comments about the Gibbs sampler are given in Appendix B which is available online.

We consider test items incorporating 8 subtasks, either required or not, to solve a specific item. Consequently, there are 256 ($= 2^8$) distinct combinations of item–feature. Obviously, the item bank from which the items are selected can include (infinitely) many more test items. All feature effects are assumed random. Knowing that the RWLLTM also includes a random intercept, the mean effects vector $\boldsymbol{\beta}$ and the individual effects vectors $\boldsymbol{\theta}_i$ are thus nine-dimensional. The true values for $\boldsymbol{\beta}$, $\mathbf{D}$ and all $\boldsymbol{\theta}_i$ were simulated and used, in turn, to simulate the responses for 1000 individuals to the selected items in their tests. The coefficients in $\boldsymbol{\beta}$ were randomly drawn from $\mathcal{U}[-2, 2]$; the variance covariance matrix $\mathbf{D}$ was computed from a simulated Cholesky factor. The diagonal elements of $\mathbf{D}$ are between 0.223 and 3.135, the off-diagonals between -0.943 and 1.323. Each vector $\boldsymbol{\theta}_i$ was then sampled from $\mathcal{N}(\mathbf{0}, \mathbf{D})$.

In the item selection criteria, the prior value $\boldsymbol{\beta}_0$ was set to the zero vector, and that of $\mathbf{D}_0$, to the identity matrix. This corresponds to very uninformative prior information, which makes

sense as in general one has little knowledge about the true effects of the features on the difficulty of the items. The integrals in the design criteria were approximated with 1024 draws. For each individual, a test design with 100 items was generated. As there are nine individual-specific coefficients in the model, a large amount of test data is necessary to obtain sufficiently accurate estimates. Unfortunately, lengthy tests may induce learning or fatigue effects. In this study, however, we assume that no item position effects are present, so that adaptive testing is warranted (Hohensinn et al., 2008; Kubinger, 2008). Response simulation and model estimation were repeated 50 times; the results reported in this section are averages over these repetitions.

From a smaller simulation study, including only 300 test takers and 30 items in their tests, it already became clear that $E\mathcal{D}_B$ is too complex, and therefore too slow, as an item selection rule for the RWLLTM. The computation time needed to select an additional item in a test is much greater than that needed by $\mathcal{D}_B$, KLP, or MUI, making $E\mathcal{D}_B$ impractical for the RWLLTM. Moreover, $E\mathcal{D}_B$ did not yield more accurate estimates for the parameters in the model. Therefore, this design criterion was discarded from the larger study and the results are only reported for $\mathcal{D}_B$, KLP and MUI. Some findings from this preliminary simulation study are given in Appendix C which is available online.

Because of the importance of random effects in the RWLLTM, we start by discussing the accuracy of the estimates for the individual coefficients $\boldsymbol{\theta}_i$. To observe the change in estimation accuracy, the RWLLTM was not only estimated with all 100 items for each individual, but also using only the first 50, 60, 70, 80 and 90 test items. Figure 1 shows the mean $\text{RMSE}_{\boldsymbol{\theta}}$ values for each design criterion and each test length. Note that the mean $\text{RMSE}_{\boldsymbol{\theta}}$ values corresponding to a random selection of test items are also added to the plot. Completely in line with expectations, a decrease in $\text{RMSE}_{\boldsymbol{\theta}}$, and therefore an increase in the estimation accuracy for the individual coefficients, is observed for increasing test lengths. Further, it is clear, and again as expected, that constructing individual tests with the $\mathcal{D}_B$, KLP or MUI criteria is much more efficient than randomly selecting the items. An intercriterion comparison between the efficient item selection algorithms on the other hand does not reveal any substantial differences: no criterion stands out in efficiency or performs significantly better than the remaining criteria.

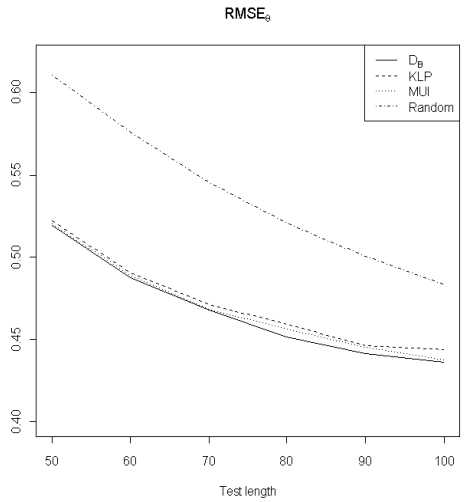

Figure 1: Mean RMSE$_{\boldsymbol{\theta}}$ values obtained with $\mathcal{D}_B$, KLP, MUI and random item selection and different test lengths

In addition to Figure 1, Figure 2 plots the average estimates of an element from $\boldsymbol{\theta}_i$ against the true values of the coefficients for all individuals obtained with either $\mathcal{D}_B$, KLP or MUI and test lengths 50 and 100. Note that the values in Figure 2 correspond to the eighth parameter. The plots for the remaining parameters are similar and therefore not displayed. The estimates obtained with the different item selection criteria are very alike, again illustrating the equivalence in design efficiency and estimation accuracy between the design algorithms.

The plots in Figure 2, however, also uncover some shrinkage of the estimated coefficients to the mean of zero. To estimate the random effects of an examinee not only the test data from that specific individual is used, but data and information from the entire population is incorporated. Estimates for the individual effects are thus, especially when there is not much individual data, shrunk to the population mean. Nevertheless, even with only 50 items in the tests, the individual coefficients are being estimated quite accurately: the correlation between the estimates and the true values, averaged over the nine parameters, is 0.9696 for $\mathcal{D}_B$, 0.9693 for KLP, and 0.9690 for MUI. Obviously, as already observed in Figure 1, the estimation of the individual feature effects is improved as the test length grows. Less shrinkage is observed in the right panel of Figure 2. Moreover, with 100 test items, the average correlations between the estimates and the true values of the individual coefficients increase to 0.9834, 0.9834, and

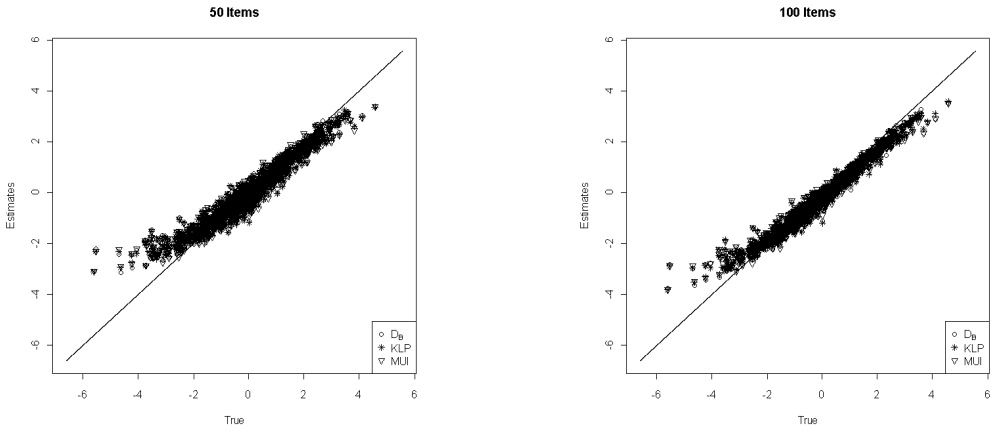

Figure 2: Estimated parameters against the true values of θ_{i8} for all individuals obtained with $\mathcal{D}_B$, KLP and MUI and test lengths 50 and 100

0.9835.

Although the plots of the estimated coefficients do not reveal any severe bias, it might be worthwhile to compare the different design criteria as to their average bias obtained over all parameters and individuals. The measure applied for this is

$$\text{BIAS} = \sum_{i=1}^{N} \sum_{m=1}^{q} \frac{|\hat{\theta}_{im} - \theta_{im}|}{Nq}. \tag{34}$$

The mean BIAS values over the 50 simulation repetitions are given in Table 2 for each design criterion and each test length (standard deviations are given between brackets). As before, so too with the bias, it appears that designing individual tests for the RWLLTM with either $\mathcal{D}_B$, KLP or MUI is much more efficient than randomly selecting the items. Between $\mathcal{D}_B$, KLP and MUI, however, no significant differences in average bias are observed.

But not only in the estimation of the individual coefficients, also for the estimation of the population parameters $\boldsymbol{\beta}$ and $\mathbf{D}$ in the RWLLTM, the design criteria perform equally efficiently. Figure 3 shows boxplots of the $\text{RMSE}_{\boldsymbol{\beta}}$ and $\text{RMSE}_{\mathbf{D}}$ values for each design criterion when the tests have 100 items. It is clear that the average estimation errors of the alternative design criteria are not significantly different. Similar conclusions can be drawn for other test lengths.

In addition to the comparison of the design efficiency of the criteria, we will have a closer look at the similarities and differences in the items they select by means of the item exposure

Table 2: *Mean BIAS Values for $\mathcal{D}_B$, KLP, MUI and Random Item Selection and Different Test Lengths (Standard Deviations in Parentheses)*

	$\mathcal{D}_B$	KLP	MUI	Random
50 items	0.388 (0.004)	0.390 (0.006)	0.388 (0.006)	0.454 (0.005)
60 items	0.362 (0.004)	0.364 (0.006)	0.363 (0.006)	0.428 (0.004)
70 items	0.346 (0.007)	0.349 (0.008)	0.346 (0.007)	0.405 (0.004)
80 items	0.333 (0.007)	0.339 (0.010)	0.337 (0.009)	0.386 (0.004)
90 items	0.326 (0.011)	0.329 (0.010)	0.328 (0.011)	0.371 (0.005)
100 items	0.322 (0.012)	0.328 (0.015)	0.323 (0.015)	0.357 (0.005)

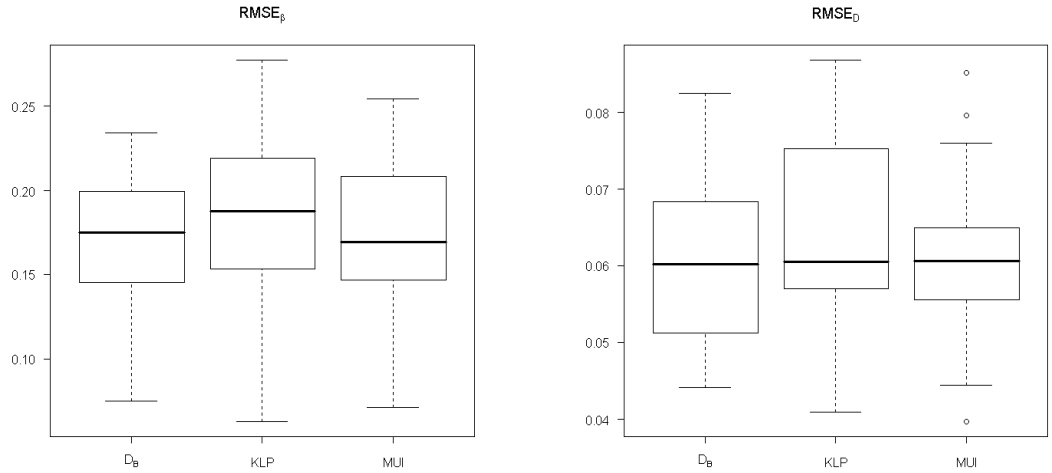

Figure 3: Boxplots of the RMSE$_{\boldsymbol{\beta}}$ and RMSE$_{\mathbf{D}}$ values obtained with $\mathcal{D}_B$, KLP and MUI and test length 100

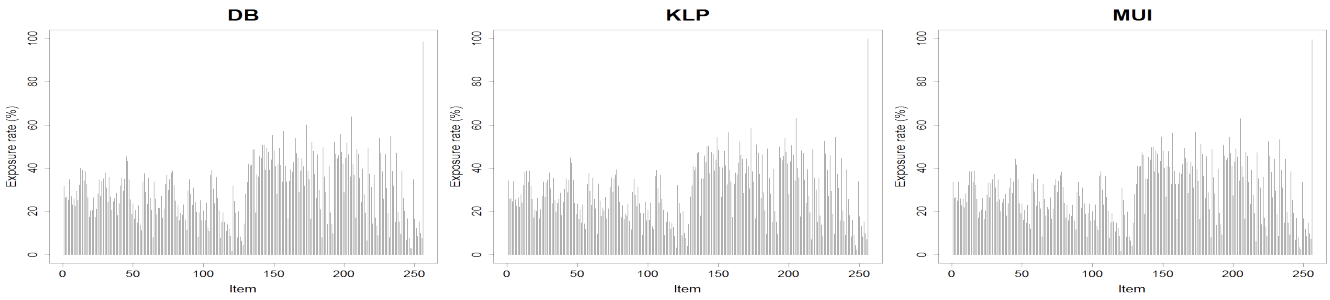

Figure 4: Exposure rates (%) for $\mathcal{D}_B$, KLP and MUI and test length 100

rates and an item overlap analysis. An exposure rate is computed for each distinct item (more specifically, each distinct item–feature combination) in the item bank, and is simply the percentage of test takers administered with that specific test item (Li and Schafer, 2005). They are given in Figure 4 for $\mathcal{D}_B$, KLP and MUI and tests with 100 items. The exposure plots are similar in shape for the other test lengths and so are not shown here. Some statistics of the exposure distributions are given in Table 3 (Li and Schafer, 2005). E.g., using the KLP design algorithm, 7.03% of the 256 items have an exposure rate greater than or equal to 5%, and strictly smaller than 10%. The exposure distribution observed with random item selection is also shown, for comparison.

There are significant similarities in the exposure rates across the criteria. All three item selection rules apply all test items over the 1000 individualized designs. Due to differences in the (high dimensional) individual feature effects, it makes sense that an accurate estimation of the many parameters in the RWLLTM requires a large variety of test items. Except for the last item, the distributions have no disturbing amounts of underexposed or overexposed items. If a test item is selected and administered too often, its content and solution become known by the test takers. Correctly solving such items therefore no longer expresses an individual's true ability, and its use could induce estimation bias. In this case, the design algorithms should be adapted to take into account item exposure control (Veldkamp and van der Linden, 2002). Underexposed items, on the other hand, may unnecessarily increase the cost of constructing the item bank.

But even if two criteria select approximately the same test items over all designs, the items

Table 3: *Item Exposure Distribution (%) for $\mathcal{D}_B$, KLP, MUI and Random Item Selection and Test Length 100*

Exposure rates	$\mathcal{D}_B$	KLP	MUI	Random
0	0.00	0.00	0.00	0.00
<0 – <5	1.56	1.56	1.56	0.00
5 – <10	6.64	7.03	7.03	0.00
10 – <15	6.25	5.08	7.03	0.00
15 – <20	13.28	14.06	13.28	0.00
20 – <25	12.11	14.06	12.11	0.00
25 – <30	13.67	12.11	12.89	0.00
30 – <35	11.33	11.72	13.28	100.00
35 – <40	13.28	14.06	12.11	0.00
40 – <45	7.81	6.64	8.59	0.00
45 – <50	8.59	8.20	7.81	0.00
50 – <60	4.30	4.69	3.52	0.00
60 – <70	0.78	0.39	0.39	0.00
70 – <80	0.00	0.00	0.00	0.00
80 – <90	0.00	0.00	0.00	0.00
90 – <100	0.39	0.00	0.39	0.00
100	0.00	0.39	0.00	0.00
Mean	29.48	29.11	28.85	32.39
Standard deviation	13.82	13.42	13.43	0.20
Minimum	1.58	2.28	1.68	31.85
Maximum	98.82	100.00	99.60	32.94

Table 4: *Item Overlap (%) for Test Lengths 50, 70 and 100*

	50 items	
	KLP	MUI
$\mathcal{D}_B$	34.51	34.10
KLP		34.91

	70 items	
	KLP	MUI
$\mathcal{D}_B$	39.06	38.64
KLP		39.19

	100 items	
	KLP	MUI
$\mathcal{D}_B$	45.39	44.95
KLP		45.23

assigned to a specific individual can of course differ. To verify whether, and if so, to what extent, the design criteria select the same items for the same test takers, an item overlap analysis was conducted. The overlap rate between two design algorithms is defined as the ratio of common items to the total amount of items in the test, averaged over individuals (Chen et al., 2000; Wang and Chang, 2011; Wang et al., 2011). The overlap rates between $\mathcal{D}_B$, KLP and MUI are given in Table 4 for test lengths of 50, 70 and 100, and appear to be fairly high. For tests with 50 items, already more than one-third of the test items are, on average, the same across the design criteria. The overlap is even approximately 45% for a test with 100 items.

From the simulations above, it can be concluded that $\mathcal{D}_B$, KLP and MUI are equally efficient as item selection rules at obtaining test data and at estimating the RWLLTM, as they have the same estimation accuracy. KLP and MUI, are, however, far less complex than the $\mathcal{D}_B$ criterion. This last requires computing the determinant of the Fisher information matrix, which greatly slows down the algorithm. For the simulation setup above, it takes, for tests of length 5, on average a little bit more than two seconds with the $\mathcal{D}_B$ criterion to select an additional item (Table 5). Although this seems acceptable, the difference from KLP and MUI is huge. Their computation time is about 0.245 seconds. Moreover, for all design criteria, the run times increase as the test progresses because more and more data must be processed to select the next item. This is clear from Table 5: the average time to generate an additional item increases with the length of the test. As the selection of an additional item in a test takes on average almost 10 times longer with $\mathcal{D}_B$ than with KLP and MUI, the latter are clearly more practical.

Note, however, that the speed of the design process is also affected by the number of candidate items, which depends on the number of item features considered. The more fea-

Table 5: *Average Computation Time (seconds) to Select a Test Item with $\mathcal{D}_B$, KLP and MUI*

Number of items	5	10	15	20
$\mathcal{D}_B$	2.331	2.402	2.460	2.489
KLP	0.241	0.253	0.256	0.266
MUI	0.248	0.262	0.269	0.276

tures, the more item–feature combinations (and thus the more candidate items to evaluate). More specifically, the number of item–feature combinations doubles for every additional binary feature. E.g., with 15 features, already 32,768 feature combinations exist, blowing up the computation time to select an item (in a test with 5 items in total) to 25 seconds even with the KLP criterion. Clearly, one should be aware of the computational cost that comes with a more complex setup of the test items.

4 Conclusion

The random weights linear logistic test model (RWLLTM) extends Fischer's linear logistic test model by incorporating individual effects for the item features. Therefore, this paper has proposed an individualized design approach for the RWLLTM. Four item selection rules were discussed and compared: the minimum posterior weighted $\mathcal{D}$-error ($\mathcal{D}_B$), the minimum expected posterior weighted $\mathcal{D}$-error ($E\mathcal{D}_B$), the maximum expected Kullback–Leibler divergence between subsequent posteriors (KLP), and the maximum mutual information (MUI). Due to an excessive complexity when being applied to the RWLLTM, the $E\mathcal{D}_B$ criterion was discarded from the simulation study.

The study clearly confirms the positive effect on the estimation accuracy from using efficiently designed tests with $\mathcal{D}_B$, KLP and MUI rather than randomly selecting the test items to estimate the RWLLTM. The results, however, do not reveal any significant differences in estimation accuracy between the different efficient design criteria. Both the individual-specific coefficients and the population parameters in the RWLLTM are estimated equally well by these algorithms. $\mathcal{D}_B$, KLP and MUI thus all appear useful in constructing efficient individualized

tests for the RWLLTM. The Kullback–Leibler criteria are, however, to be given the preference due to their speed. The computation of KLP and MUI is far less complex, causing a huge decrease in the computation time for an additional item in a test, in comparison to $\mathcal{D}_B$. The former criteria make individualized test design for the RWLLTM feasible.

Acknowledgements

Research funded by ZKC1090 / DBOF/08/014 - DBOF project of the KU Leuven

References

Allenby, G.M., Arora, N., Ginter, J.L., 1995. Incorporating prior knowledge into the analysis of conjoint studies. Journal of Marketing Research. 32, 152–162.

Arora, N., Allenby, G.M., Ginter, J.L., 1998. A hierarchical Bayes model of primary and secondary demand. Marketing Science. 17, 29–44.

Arora, N., Huber, J., 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. Journal of Consumer Research. 28, 273–283.

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Clarendon Press, Oxford.

Bouwmeester, S., van Rijen, E.H.M., Sijtsma, K., 2011. Understanding phoneme segmentation performance by analyzing abilities and word properties. European Journal of Psychological Assessment. 27, 95–102.

Chang, H.H., Ying, Z., 1996. A global information approach to computerized adaptive testing. Applied Psychological Measurement. 20, 213–229.

Chen, S.Y., Ankenmann, R.D., Chang, H.H., 2000. A comparison of item selection rules at the early stages of computerized adaptive testing. Applied Psychological Measurement. 24, 241–255.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., Partchev, I., 2011. The estimation of item response models with the `lmer` function from the lme4 package in R. Journal of Statistical Software. 39, 1–28.

Embretson, S.E., Daniel, R.C., 2008. Understanding and quantifying cognitive complexity level in mathematical problem solving items. Psychology Science Quarterly. 50, 328–344.

Fischer, G.H., 1973. The linear logistic test model as an instrument in educational research. Acta Psychologica. 37, 359–374.

Freund, P.A., Hofer, S., Holling, H., 2008. Explaining and controlling for the psychometric properties of computer-generated figural matrix items. Applied Psychological Measurement. 32, 195–210.

Hohensinn, C., Kubinger, K.D., Reif, M., Holocher-Ertl, S., Khorramdel, L., Frebort, M., 2008. Examining item-position effects in large-scale assessment using the linear logistic test model. Psychology Science Quarterly. 50, 391–402.

Holling, H., Bertling, J.P., Zeuch, N., 2009. Automatic item generation of probability word problems. Studies in Educational Evaluation. 35, 71–76.

Hornke, L.F., Habon, M.W., 1986. Rule-based item bank construction and evaluation

within the linear logistic framework. Applied Psychological Measurement. 10, 369–380.

Kubinger, K.D., 2008. On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. Psychology Science Quarterly. 50, 311–327.

Li, Y.H., Schafer, W.D., 2005. Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. Applied Psychological Measurement. 29, 3–25.

Medina-Diaz, M., 1993. Analysis of cognitive structure using the linear logistic test model and quadratic assignment. Applied Psychological Measurement. 17, 117–130.

Mulder, J., van der Linden, W.J., 2009. Multidimensional adaptive testing with optimal design criteria for item selection. Psychometrika. 74, 273–296.

Mulder, J., van der Linden, W.J., 2010. Multidimensional adaptive testing with Kullback–Leibler information item selection. In: van der Linden, W.J., Glas, C.A.W. (eds.) Elements of Adaptive Testing. pp. 77–101. Springer-Verlag, New York.

Poinstingl, H., 2009. The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. Psychology Science Quarterly. 51, 123–134.

Rijmen, F., De Boeck, P., 2002. The random weights linear logistic test model. Applied Psychological Measurement. 26, 271–285.

Rupp, A.A., Templin, J., Henson, R.A., 2010. Diagnostic Measurement. Theory, Methods,

& Applications. New York: Guilford Press.

Segall, D.O., 2010. Principles of multidimensional adaptive testing. In: van der Linden, W.J., Glas, C.A.W. (eds.) Elements of Adaptive Testing. pp. 57–75. Springer-Verlag, New York.

van der Linden, W.J., 1998. Bayesian item selection criteria for adaptive testing. Psychometrika. 63, 201–216.

van der Linden, W.J., Pashley, P.J., 2010. Item selection and ability estimation in adaptive testing. In: van der Linden, W.J., Glas, C.A.W. (eds.) Elements of Adaptive Testing. pp. 3–30. Springer-Verlag, New York.

Veldkamp, B.P., van der Linden, W.J., 2002. Multidimensional adaptive testing with constraints on test content. Psychometrika. 67, 575–588.

Verbeke, G., Molenberghs, G., 2003. The use of score tests for inference on variance components. Biometrics. 59, 254–262.

Wang, C., Chang, H.H., 2011. Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. Psychometrika. 76, 363–384.

Yu, J., Goos, P., Vandebroek, M., 2011. Individually adapted sequential Bayesian conjoint-choice designs in the presence of consumer heterogeneity. International Journal of Research in Marketing. 28, 378–388.

Zeuch, N., Holling, H., Kuhn, J.T., 2011. Analysis of the latin square task with linear logistic test models. Learning and Individual Differences. 21, 629–632.